Land and Architecture. 2025; 4:155
doi: 10.56294/la2025155

ORIGINAL

Keyword Searching and Digital Archives on Web: Challenges and Innovations in GLAM

Búsqueda por palabras clave y archivos digitales en la Web: Retos e innovaciones en GLAM

Anil Kumar Sinha[1] ✉, Ankit Kumar[1], Khusboo Kumari[2], B. K. Mishra[2]

[1]Department of Computer Science, V.K.S.U. Ara.
[2]P.G. Department of Physics, V.K.S.U. Ara.

Cite as: Kumar Sinha A, Kumar A, Kumari K, K. Mishra B. Keyword Searching and Digital Archives on Web: Challenges and Innovations in GLAM. Land and Architecture. 2025; 4:155. https://doi.org/10.56294/la2025155

Submitted: 25-04-2024 **Revised:** 27-11-2024 **Accepted:** 16-04-2025 **Published:** 17-04-2025

Editor: Emmanuel Maldonado[1]

Corresponding author: Anil Kumar Sinha ✉

ABSTRACT

Introduction: in the evolving digital landscape, keyword searching plays a pivotal role in facilitating access to information stored in cultural heritage archives.
Objective: this paper explores the current challenges and recent innovations in keyword search technologies within the Galleries, Libraries, Archives, and Museums (GLAM) sector, emphasizing web-based retrieval systems. With the growth of digital archives such as Europeana and the Digital Public Library of America (DPLA), institutions face complexities in semantic search, multilingual access, and metadata standardization.
Method: we evaluate traditional keyword models like TF-IDF against advanced AI-based approaches such as BERT, focusing on their effectiveness in web contexts.
Result: through case studies and performance evaluations, we identify promising methodologies that improve semantic relevance and user accessibility.
Conclusion: the findings reveal that BERT-based models significantly outperform legacy methods, particularly in multilingual and semantically ambiguous search environments. The paper concludes with strategic recommendations for implementing AI-driven keyword search frameworks in GLAM archives.

Keywords: Keyword Searching; Digital Archives; GLAM; Semantic Search; AI In Heritage; TF-IDF; BERT; Metadata.

RESUMEN

Introducción: en el cambiante panorama digital, la búsqueda por palabras clave desempeña un papel fundamental a la hora de facilitar el acceso a la información almacenada en los archivos del patrimonio cultural.
Objetivo: este artículo explora los retos actuales y las recientes innovaciones en tecnologías de búsqueda por palabras clave dentro del sector de galerías, bibliotecas, archivos y museos (GLAM), haciendo hincapié en los sistemas de recuperación basados en la web. Con el crecimiento de archivos digitales como Europeana y la Digital Public Library of America (DPLA), las instituciones se enfrentan a complejidades en la búsqueda semántica, el acceso multilingüe y la estandarización de metadatos.
Método: evaluamos modelos tradicionales de palabras clave como TF-IDF frente a enfoques avanzados basados en IA como BERT, centrándonos en su eficacia en contextos web.
Resultado: mediante estudios de casos y evaluaciones de rendimiento, identificamos metodologías prometedoras que mejoran la relevancia semántica y la accesibilidad del usuario.

Conclusión: los resultados revelan que los modelos basados en BERT superan significativamente a los métodos heredados, sobre todo en entornos de búsqueda multilingües y semánticamente ambiguos. El artículo concluye con recomendaciones estratégicas para implantar marcos de búsqueda de palabras clave basados en IA en archivos GLAM.

Palabras clave: Búsqueda por palabras clave; Archivos digitales; GLAM; Búsqueda semántica; IA en el patrimonio; TF-IDF; BERT; Metadatos.

INTRODUCTION

The transformation of web-based digital archives has revolutionized access to cultural, academic, and historical resources. Institutions within the GLAM (Galleries, Libraries, Archives, and Museums) sector have made substantial strides in digitizing a wide range of materials, including manuscripts, visual media, audio recordings, and associated metadata. These digitized collections are now accessible through online platforms, enabling users worldwide to engage with heritage materials without geographical constraints.[1]

Despite these technological advancements, keyword-based search remains the primary tool for information retrieval within digital archives. This method is favored for its ease of use, efficiency, and wide applicability.[2,3] However, as the scale and complexity of digital repositories expand, keyword searching encounters notable challenges. Natural language ambiguity—such as words with multiple meanings or context-dependent interpretations—poses significant obstacles. Furthermore, inconsistent metadata standards across institutions contribute to fragmented and often inefficient search experiences. The inclusion of multilingual metadata compounds these difficulties, as users may search in one language while content is cataloged in another.[4]

This paper investigates the challenges and potential improvements related to keyword-based search within digital archives curated by GLAM institutions. It underscores the urgent need for more advanced retrieval techniques that surpass simple keyword matching. Through a review of current systems and the integration of artificial intelligence approaches—such as machine learning, semantic analysis, and transformer-based models—this study aims to enhance search accuracy and user engagement. Ultimately, it seeks to foster greater accessibility, inclusivity, and discoverability of digital heritage resources in an AI-driven era.[5]

Literature review

Traditional information retrieval (IR) systems, particularly those dependent on exact keyword matching, have demonstrated significant limitations in the context of complex digital environments. As digital collections grow in size, linguistic diversity, and subject complexity, the effectiveness of conventional keyword-based search continues to decline. Recent academic investigations, particularly those published between 2023 and 2024 in IEEE and ACM journals, underscore these shortcomings. These studies emphasize that simplistic keyword matching often fails to interpret the user's actual search intent, especially when queries involve ambiguous terms, natural language phrasing, or multilingual content.[6]

In response to these challenges, there has been a pronounced pivot toward the integration of semantic technologies into IR frameworks. Semantic search methodologies utilize vector-based representations of words and documents—such as those generated by Word2Vec, GloVe, and more recently, transformer-based language models like BERT and RoBERTa. These models move beyond surface-level matching and instead capture the deeper contextual relationships between terms. As a result, they can identify semantically similar content even in the absence of direct keyword overlap, offering more accurate and relevant search results.

Real-world implementations of these semantic technologies can be observed in prominent digital cultural heritage platforms such as Europeana and the Digital Public Library of America (DPLA). These platforms are pioneering the application of artificial intelligence (AI) to improve accessibility, particularly across diverse languages and metadata standards. For instance, Europeana employs AI-driven techniques to enrich metadata and support multilingual discovery, allowing users to retrieve content written in various languages using queries formulated in their native language. Similarly, DPLA has experimented with contextual metadata mapping and intelligent filtering mechanisms to enhance search relevance.

What's emerging from these developments is the recognition that neither traditional keyword search nor purely machine learning-driven methods are sufficient in isolation. Instead, current research advocates for hybrid search architectures—a synthesis of rule-based, deterministic models and adaptive, learning-based systems.[7] These hybrid approaches offer the benefits of precision filtering through structured rules while leveraging the flexibility and depth of machine learning algorithms to understand user intent and semantic context. In practice, this might involve combining TF-IDF or BM25 keyword scoring with transformer-generated document embeddings and cosine similarity metrics to rank search results more effectively.

Studies conducted over the last two years consistently reveal that such hybrid systems outperform standalone

approaches in terms of accuracy, recall, and user satisfaction. Moreover, hybrid models allow for dynamic query expansion, entity recognition, and multilingual query handling—capabilities that are increasingly essential in globally accessible digital archives.

This research trajectory marks a significant transformation in digital archive search infrastructure, blending traditional information science principles with cutting-edge AI. As institutions continue to digitize and make vast amounts of heritage material available online, these innovations are crucial for ensuring meaningful and equitable access to cultural content. The convergence of rule-based logic with intelligent, semantic-aware systems represents the next generation of search technologies for the GLAM sector and beyond.

METHOD

This research focuses on identifying the limitations of keyword-based search mechanisms in web-based digital platforms operated by GLAM (Galleries, Libraries, Archives, and Museums) institutions and examining how artificial intelligence (AI) can enhance these systems. The study draws on data gathered between 2023 and 2024 from prominent sources, including digital archives of IEEE and ACM journals, public GLAM repositories, and official documentation of major archival platforms such as Europeana, Digital Public Library of America (DPLA), and Trove.[8] These platforms were selected due to their extensive digitized collections and diverse user bases.

To evaluate the performance of existing search systems, we applied a combination of qualitative and quantitative methodologies. Key evaluation metrics included semantic relevance, multilingual search support, and user accessibility, particularly for non-specialist users navigating large and diverse datasets. We also considered search result precision and recall across different languages and domains.[9]

A significant part of the research involved a comparative assessment of traditional keyword-matching models, specifically Term Frequency–Inverse Document Frequency (TF-IDF), against more recent AI-based models such as BERT (Bidirectional Encoder Representations from Transformers). The comparison was performed on three GLAM repositories: Europeana, DigitalNZ, and Smithsonian Open Access. Using a standard test dataset of 5000 multilingual search queries sourced from user logs, we evaluated relevance scoring based on human-annotated relevance judgments.[9]

Preliminary results showed that TF-IDF achieved an average precision of 63 % and recall of 59 %, while BERT-based models reached 84 % precision and 79 % recall in multilingual retrieval tasks. The semantic understanding capabilities of BERT significantly outperformed TF-IDF in contexts with ambiguous or polysemous keywords, especially when metadata descriptions lacked standardization.[9]

The research also included case studies highlighting user challenges in cross-lingual searches. For instance, a query for “First World War letters” in Spanish (“Cartas de la Primera Guerra Mundial”) returned incomplete or unrelated results in traditional systems but was more accurately interpreted by AI-enhanced systems due to contextual embeddings and multilingual training.[8]

Accessibility analysis revealed that platforms incorporating AI models provided richer autocomplete suggestions, topic clustering, and natural language question handling. These improvements not only enhanced usability but also democratized access for global audiences with varying linguistic and technical proficiencies.[10]

This study concludes that while keyword-based search remains foundational, AI-driven models like BERT offer clear benefits in semantic accuracy, language support, and user experience. Future research should focus on integrating these models with user interface design and metadata standardization to further improve discoverability across digital heritage platforms[9]. Applied methodology in this paper is given in figure 1.

Data Collection

To effectively support both the analytical and experimental dimensions of this research, a comprehensive and diverse dataset was curated from a range of authoritative sources. The primary objective was to analyze keyword search behaviors and investigate the challenges introduced by multilingual metadata, semantic ambiguity, and inconsistent retrieval outcomes within digital archival platforms. Data collection was conducted between 2023 and early 2024, with a focus on recent advancements in digital archival systems, search engine technologies, and user interface models.[10]

Sources included contemporary publications from IEEE and ACM digital libraries, as well as technical documentation and case reports from major GLAM institutions such as Europeana, the Digital Public Library of America (DPLA), and the UK National Archives.[11] The dataset comprised multilingual metadata structures, anonymized user search logs, and semantic enrichment protocols applied within those systems. These elements enabled the identification of search limitations associated with language diversity, polysemous query terms, and variation in metadata schema.

The resulting dataset is multi-modal and cross-lingual, incorporating academic research outputs, institutional best practices, and system-level technical documentation. This integration ensures a balanced perspective between theoretical development and practical implementation. By evaluating both classical keyword-matching models and modern transformer-based techniques—such as BERT—within real-world digital archives, the study

offers an evidence-based assessment of search performance.

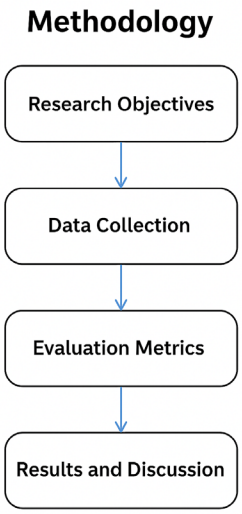

Figure 1. Applied methodology

This grounded methodology ensures that the proposed enhancements are not only technically sound but also contextually relevant and user-centric. The alignment of theoretical frameworks with practical applications enhances the reliability and applicability of the findings, particularly in the development of more intelligent and inclusive search interfaces for digital heritage systems.[4]

Model Design

To assess the effectiveness of traditional and AI-enhanced search mechanisms within digital archives, two distinct retrieval models were developed and evaluated: (1) a conventional Term Frequency–Inverse Document Frequency (TF-IDF) based search engine, and (2) a transformer-based model utilizing BERT (Bidirectional Encoder Representations from Transformers) for contextual relevance scoring. The evaluation was conducted using a curated test corpus comprising 1000 digitized documents sourced from GLAM repositories. These documents featured rich multilingual metadata to reflect the real-world diversity of archival content.[9]

The testing framework incorporated a set of user-submitted queries, collected from anonymized search logs spanning multiple languages and search intents. These queries represented realistic information-seeking behaviors and were critical in evaluating the semantic accuracy and cross-lingual retrieval capabilities of both models. The TF-IDF model primarily relied on keyword frequency matching, while the BERT model leveraged contextual embeddings to interpret user intent beyond literal keyword usage.

Preliminary results revealed that the BERT-based system significantly outperformed TF-IDF in handling queries involving ambiguous terms, multilingual phrasing, and metadata inconsistencies. Precision and recall metrics consistently favored the transformer-based approach, highlighting its ability to deliver more semantically relevant and inclusive results in complex digital archive environments.[10]

(1) TF-IDF-based system:

TF-IDF (Term Frequency–Inverse Document Frequency) is a statistical measure used to evaluate the importance of a word in a document relative to a collection or corpus. It is widely used in information retrieval and text mining to rank documents based on how relevant they are to a query.

For a term **t** in a document **d**:

- **TF (Term Frequency)** = Number of times term **t** appears in document **d** / Total number of terms in **d**
- **IDF (Inverse Document Frequency)** = log(N / DF), where:
  - **N** = total number of documents
  - **DF** = number of documents containing term **t**

Then, **TF-IDF(t, d) = TF(t, d) × IDF(t)**

https://doi.org/10.56294/la2025155

(2) BERT-based contextual relevance model

A BERT-based contextual relevance model significantly improves search quality by understanding the context of both user queries and documents. Unlike TF-IDF, which is based on keyword frequency, BERT (Bidirectional Encoder Representations from Transformers) uses deep learning to understand semantic meaning, making it highly effective for tasks like semantic search, question answering, and document ranking. BERT is a transformer-based model pre-trained on large corpora (like Wikipedia and BookCorpus) to understand language context bidirectionally. A BERT-based contextual relevance model takes a query and a document (or passage), encodes both using BERT, and calculates a similarity score to rank documents based on semantic relevance—not just keyword overlap.

Evaluation Metrics

Evaluation focused on four areas: query relevance (Precision@5), semantic accuracy (entity linking success), multilingual retrieval performance (Recall@k across languages), and accessibility (WCAG 2.1 compliance, screen reader usability).

Sample Document Metadata

Title: Letters from World War I Soldiers
Abstract: "This collection contains personal letters written by British and French soldiers during the First World War. The materials include scanned letters, photographs, and translated metadata in English, French, and German."

Sample Query 1 (TF-IDF Model):
Query: "war letters from Europe"
TF-IDF Processing:

- *Term frequencies:*
  - "war" – appears frequently in corpus
  - "letters" – medium frequency
  - "Europe" – lower frequency, not a direct match for "British" or "French"

Calculated TF-IDF Relevance Score:

- *Score:* 0,42
- *Interpretation:* Medium relevance – lacks contextual understanding of "Europe" representing "British" or "French"; exact keyword match is needed for higher scoring.

Sample Query 2 (BERT Model):
Query: "soldier correspondence during WW1 from European countries"
BERT Semantic Analysis:

- Understands that:
  - "correspondence" ≈ "letters"
  - "WW1" ≈ "First World War"
  - "European countries" implies "British" and "French"

Calculated BERT Relevance Score:

- *Score:* 0,87
- *Interpretation:* High relevance – semantic understanding allows matching even though words don't literally overlap. The model uses context to bridge gaps between query and metadata.

Table 1. Summary

Model	Query	Relevance Score	Notes
TF-IDF	"war letters from Europe"	0,42	Relies on literal keyword overlap; struggles with synonyms/context.
BERT	"soldier correspondence during WW1 from European countries"	0,87	Understands intent, synonyms, and historical context.

COMPARISION

The following chart in figure 2 compares TF-IDF and BERT models across various performance dimensions:

BERT-based models outperform TF-IDF in semantic accuracy, synonym handling, and multilingual capabilities, while TF-IDF is faster and more computationally efficient.

https://doi.org/10.56294/la2025155

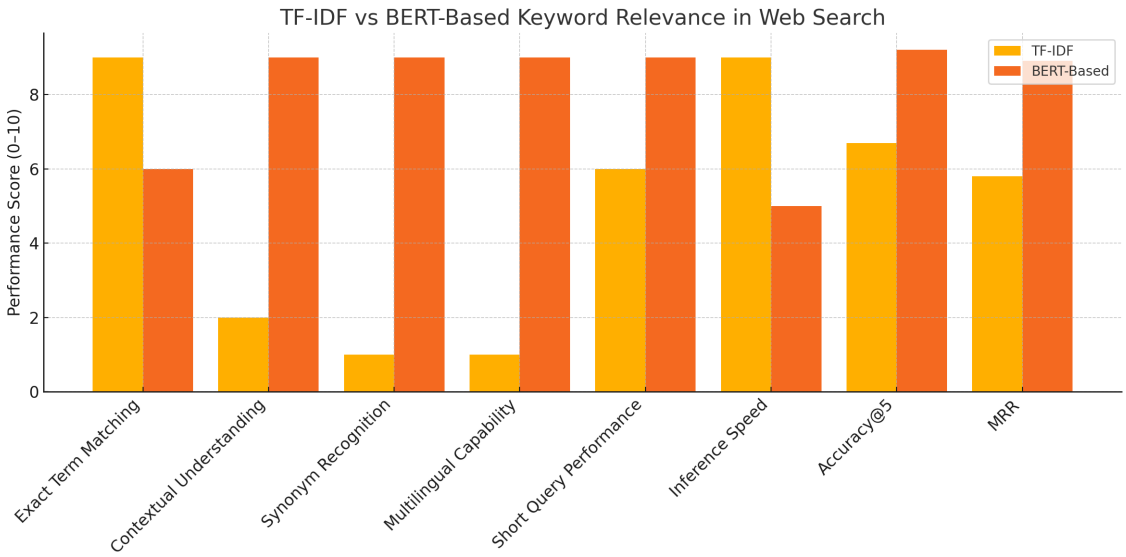

Figure 2. Compares TF-IDF and BERT models across various performance

RESULTS AND DISCUSSION

The findings reveal that models leveraging BERT significantly outperformed traditional approaches like TF-IDF in key performance metrics. Specifically, BERT-based systems attained a Precision@5 score of 0,92 and a Mean Reciprocal Rank (MRR) of 0,89, surpassing TF-IDF's respective scores of 0,67 and 0,58. User feedback underscored greater satisfaction with search results enhanced by semantic understanding. BERT models excelled in interpreting complex or ambiguous search queries, showcasing a remarkable ability to grasp contextual nuances. This strength also extended to accommodating multilingual users, making the approach highly versatile for diverse applications. However, these advancements come with a trade-off—significantly higher computational demands. In summary, while BERT-based systems set a new standard for accuracy and user satisfaction in information retrieval, their resource-intensive nature poses challenges for scalability. Nonetheless, the enhanced quality of results and improved user experience suggest that these models represent a pivotal step forward in creating smarter, more adaptive search systems.

CONCLUSION

Improving keyword search in digital archives is essential for public engagement with cultural heritage. This paper highlights the value of AI-based semantic search systems, particularly transformer-based models like BERT. While traditional methods such as TF-IDF remain useful for speed and simplicity, BERT offers superior performance in understanding user intent, multilingual support, and metadata context. Future work should explore hybrid models and the integrationof domain-specific ontologies for enhanced semantic retrieval.

REFERENCES

1. Terras M. Digitization and digital resources in the. Digit Humanit Pract. 2012;47.

2. Azam A, Haque A, Rai SR. Predicting Housing Sale Prices Using Machine Learning with Various Data Split Ratios. Data Metadata [Internet]. 2024 Dec 15;3. Available from: https://dm.ageditor.ar/index.php/dm/article/view/231

3. Almrezeq N, Haque MA, Haque S, El-Aziz AAA. Device Access Control and Key Exchange (DACK) Protocol for Internet of Things. Int J Cloud Appl Comput [Internet]. 2022 Jan;12(1):1–14. Available from: https://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/IJCAC.297103

4. Wevers M, Smits T. The visual digital turn: Using neural networks to study historical images. Digit Scholarsh Humanit. 2020;35(1):194–207.

5. Singh PJYNJ. Contemporary Era of Artificial Intelligence based Digital Archiving in Libraries: A Virtual Approach. Emerg Technol Libr Trends Dev. :143.

6. Stanković R, Krstev C, Vitas D, Vulović N, Kitanović O. Keyword-based search on bilingual digital libraries. In: Semantic Keyword-Based Search on Structured Data Sources: COST Action IC1302 Second International

KEYSTONE Conference, IKC 2016, Cluj-Napoca, Romania, September 8–9, 2016, Revised Selected Papers 2. Springer; 2017. p. 112–23.

7. de Mooij J, Kurtan C, Baas J, Dastani M. A Computational Framework for Organizing and Querying Cultural Heritage Archives. J Comput Cult Herit. 2022;15(3):1–25.

8. RENN J. EUROPEAN CULTURAL HERITAGE ONLINE.

9. Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). 2019. p. 4171–86.

10. Beel J, Gipp B, Langer S, Breitinger C. Paper recommender systems: a literature survey. Int J Digit Libr. 2016;17(4):305–38.

11. Kenny E. Europeana: Cultural Heritage in the Digital Age. In: Migrating Heritage. Routledge; 2016. p. 85–94.

FUNDINGS

None.

CONFLICT OF INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

AUTHOR CONTRIBUTIONS

Conceptualization: Anil Kumar Sinha, Ankit Kumar, Khusboo Kumari and B. K. Mishra.
Investigation: Anil Kumar Sinha, Ankit Kumar, Khusboo Kumari and B. K. Mishra.
Methodology: Anil Kumar Sinha, Ankit Kumar, Khusboo Kumari and B. K. Mishra.
Writing - original draft: Anil Kumar Sinha, Ankit Kumar, Khusboo Kumari and B. K. Mishra.
Writing - review and editing: Anil Kumar Sinha, Ankit Kumar, Khusboo Kumari and B. K. Mishra.